DOCUMENT RESUME

ED 465 800 TM 034 187

AUTHOR Romano, Jeanine; Kromrey, Jeffrey D.
TITLE The "RG Sausage's" Missing Ingredients: Investigating the Validity of Reliability Generalization Study Design.
PUB DATE 2002-04-00
NOTE 77p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Error of Measurement; Monte Carlo Methods; Reliability; *Research Methodology; Scores; Simulation; *Validity

ABSTRACT

The purpose of this study was to examine the potential impact of selected methodological factors on the validity of conclusions from reliability generalization (RG) studies. The study focused on four factors; (1) missing data in the primary studies; (2) transformation of sample reliability estimates; (3) use of sample weights for estimating mean score reliability and building confidence bands; and (4) differences between analyses of score reliability estimates and estimates of standard error of measurement. The research was a Monte Carlo study in which random samples were simulated under known and controlled population conditions. In the Monte Carlo study, RG studies were simulated by generating samples in primary studies, estimating the reliability of scores in these samples, and then aggregating the sample reliability estimates in the RG studies. In general, the results suggest that the use of Fishers z transformation of the reliability estimates provided a modest increase in the accuracy of the estimation of the population mean reliability. Although the statistical bias in point estimates of the mean reliability were very small across most conditions, the confidence bands obtained using the Fisher transformation were more accurate than the confidence bands obtained using the untransformed r. To refer to a metaphor developed by B. Thompson and Y. Vacha-Haase (2000), the RG chef needs to make sure that the ingredients are measured and prepared correctly to ensure that the RG sausage does not leave the reader with an unpleasant aftertaste. (Contains 6 figures, 23 tables, and 27 references.) (SLD)

Running head: Reliability Generalization

The "RG Sausage's" Missing Ingredients:

Investigating the Validity of Reliability Generalization Study Design

Jeanine Romano

University of Tampa

Jeffrey D. Kromrey

University of South Florida

Address Correspondence to:

Jeanine Romano
Dept. of Mathematics
University of Tampa
401 W. Kennedy Blvd, SC 254
Tampa, FL 33606
(813) 253-3333 x 3122
romano@ut.edu

Paper presented at the annual meeting of the American Educational Research Association, April 1 – 5, 2002, New Orleans,LA

**Validity of Reliability Generalization Study Design:
Examining the "RG Sausage's" Missing Ingredients**

In 1998 a meta-analytic method called reliability generalization (RG) was proposed by Vacha-Hasse to describe estimated measurement error in a test's scores across studies. RG can also be used to analyze measurement error in different scales that measure the same construct. This method is similar to that used in validity generalization studies that describe the extent to which validity evidence for scores is generalizable across research contexts (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977). In RG studies the dependent variable is a statistical index of score dependability (typically, the score reliability estimate). RG studies can be used to investigate the distribution of reliability estimates across studies and to identify study characteristics which may be related to variation in reliability estimates, such as sample size, type of reliability estimate (coefficient alpha vs. test-retest), different forms of an instrument, or other special characteristics (Henson, 2001;Vacha-Haase, 1998).

The Reliability of Measures

In classical test theory, the reliability coefficient, ρ_{xx}, is defined as the correlation between scores on parallel tests (Crocker and Aligina., 1986). According to classical test theory an examinee's observed score, X, can be expressed as the sum of his/her true score and random error:

$$X = T + E$$

The reliability coefficient is the proportion of the observed variance in scores that represents true score variance rather than random error:

$$\rho_{xx} = \frac{\sigma^2_{true}}{\sigma^2_{total}}$$

where ρ_{xx} is the ratio of the true score variance to total score variance.

The most common approaches for estimating the reliability of scores include administering the same test twice to the same examinees (test-retest reliability) or administering the test only once and estimating score reliability from the intercorrelation of test items (internal consistency estimates). Test-retest reliability is estimated by calculating the correlation coefficient between the scores obtained on the two administrations of the test. Internal consistency reliability is estimated by calculating the correlations between subsets of items on the test (Crocker & Algina, 1986)

Standard Error of the Measurement (SEM)

An alternative index of score reliability is the standard error of measurement (SEM). The SEM is calculated from the sample standard deviation of observed scores, $\hat{\sigma}_x$, and the estimated reliability coefficient, r_{xx}, of a given instrument, such that SEM = $\hat{\sigma}_x \sqrt{1-r_{xx}}$ (Crocker & Algina, 1986). This statistic is an estimate of the standard deviation of observed scores around a given true score, that is, the standard deviation of the measurement errors. For example if the reliability coefficient , r_{xx}, was .75 and the standard deviation, $\hat{\sigma}_x$, on an instrument was 3.2 then the SEM = $3.2\sqrt{1-.75} = 1.6$. It is important to remember that the standard error of measurement is a function of *both* score reliability and score variability.

Meta-analysis

Meta-analysis, sometimes referred to as research synthesis, is a quantitative research design that converts individual study outcomes to a common metric, such as effect sizes, and compares them across studies. Each study is considered one observation from a hypothetical universe of studies. In 1976, Glass originated the term 'meta-analysis' and defined it as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating findings" (p.3). For Glass's meta-analysis procedure the mean effect size, $\overline{d}$, is an estimate of the population effect size, δ, across studies for the entire universe of studies. Meta-analysis is considered a secondary research method that can be used to quantitatively summarize large bodies of literature. When a large number of studies are aggregated, meta-analysis can investigate factors that were not investigated in the primary studies and detect the effect of possible moderating variables. In terms of RG, we are looking across studies at measurement error and attempting to characterize the psychometric properties of the hypothetical universe of studies that may employ a particular measure. Such properties may include the mean reliability coefficient obtained in such a population, the variance of the reliability coefficient across studies, and research design factors that may influence the magnitude of the coefficient (i.e., moderating variables).

In the aggregation of research results through meta-analysis, fundamental questions typically focus on (a) point and interval estimation of the mean effect size, and (b) the relationship between the mean effect size and research design factors. For example, studies investigating the use of computers in mathematics instruction may provide an estimate of the typical effect size of the achievement difference between students using computers and those not using computers. In addition, researchers may hypothesize a positive relationship between the amount of time spent using computer software and measures of mathematics achievement. Such estimates of mean effect sizes and relationships between effect sizes and

other variables are usually obtained using weighted least squares, in which individual studies' effect sizes are weighted by the inverse of their sampling variance (Hedges & Olkin, 1985). That is,

$$v_i = \frac{1}{\text{var}\left(\hat{\delta}_i\right)}$$

where v_i = weight for the i^{th} effect size, and

$\text{var}\left(\hat{\delta}_i\right)$ = estimated sampling variance of the i^{th} effect size.

The use of such weights in statistical estimates gives greater credence to effect sizes obtained from studies with less sampling error (typically, those based on larger sample sizes).

Methodological Issues in RG Studies

At the time of this writing, only ten RG studies have been published (see Table 1). Of these, only one (Henson, Kogan & Vacha-Hasse, 2001), has examined reliability generalization in terms of multiple measures of the same construct. Of the ten, eight included both test-retest reliability and internal consistency estimates and two examined only internal consistency estimates.

Potential methodological problems are evident in RG studies and the debate about their solution has only just begun (Sawilowsky, 2000; Thompson & Vaccha-Haase, 2000; Helms 1999). The major controversies include (a) approaches for treatment of large proportions of missing data in the published literature, (b) appropriate analyses of reliability estimates that are not statistically independent, (c) the use of nonlinear transformations of sample reliability estimates, (d) the need to weight the observed sample statistics to account for differences in sampling error across studies and (e) the differences between analyses of reliability coefficients and analyses of the estimated standard errors of measurement (SEM).

Not only are RG studies similar to meta-analytic studies in terms of methods and goals, they also are similar in terms of publication bias. In meta-analysis, publication bias is sometimes referred to as the "file-drawer problem" (Rosenthal, 1979). In most cases, meta-analyses are conducted using only published studies which may be biased towards statistically significant results. The missing data problem is exacerbated in RG studies because information on reliability often is not reported or the reported reliability estimates are based on instruments' technical manuals rather than based on the sample used in the research. Vacha-Hasse , Kogan, and Thompson (2000) refer to this as "reliability-induction" which is, "...used to refer to the practice of explicitly referencing the reliability coefficients from prior research as the sole warrant for presuming the score integrity of entirely new data" (p. 512). The tendency for published research to neglect estimates

of score reliability yields data sources with very large proportions of missing information. For example in their RG study of the Beck Depression Inventory (BDI) scores, Yin and Fan (2000), found that out of 1200 studies that used the BDI 80.1% (n = 961) did not mention reliability at all, 5.6% (n = 67) mentioned it with no citation of the estimate's source and 6.8% (n = 82) cited reliability from the published test manuals or other sources, leaving only 7.5% (n = 90) of the studies that reported reliability coefficients for the data used in the actual studies. The lack of reporting of reliability coefficients for the data in hand is an unfortunate common occurrence (Thompson & Snyder, 1998; Vacha-Hasse, Ness, Nilsson, & Reetz,1999).

Another important issue to consider is the fact that in several of the RG studies the samples analyzed did not represent independent observations. For example, Yin and Fan's (2000) RG study on the BDI included 164 reliability coefficients from 90 studies. Similarly, Vacha-Hasse's (1998) RG study on the Bem Sex Role Inventory (BSRI) used 87 reliability coefficients from 57 studies; and Caruso's (2000) RG study on the NEO personality scale used 51 reliability estimates from 37 studies. These are clearly violations of independence of observations.

In addition, some debate has been voiced in terms of using Fisher-z's transformation to normalize reliability estimates when conducting an RG study (Sawilowsky, 2000). Thompson and Vacha-Hasse (2000) have argued that reliability coefficients are a squared metric (i.e., the squared correlation between observed scores and "true" scores) and consequently the Fisher's z transformation is unnecessary.

With regards to the issue of sample weighting (i.e., weighting each reliability coefficient by an estimate of its sampling error), the use of differential weights is relatively rare in RG studies. Although Yin and Fan (2000) used a weighted analysis in their RG study, it is not common practice in this new area of research. In his RG study on the NEO personality scales, Caruso (2000) addressed this issue but argued that because the sample sizes ranged from $n = 21$ to $n = 3,856$ the large samples would have much more influence than small samples. He also stated that because he found no statistically significant correlation between sample size and reliability, sample weighting was unnecessary. Finally, he indicated that he ran an analysis using sample size weights and the results were no different than those obtained from the unweighted analysis.

Finally, interest has developed in the similarities and differences between RG analyses based on reliability coefficients and those based on SEM. For example, in their RG study on the BDI, Yin and Fan (2000) argue that the standard error should be reported because SEM is a function of both group variability and the reliability estimate. They argue that there is not an inverse relationship between the SEM and the reliability estimate, i.e. "... a lower reliability estimate does not necessarily mean the corresponding SEM will be larger" (p. 206). While Thompson and Vacha-Haase

(2000) agree that an RG study can be accomplished using the SEM, they remind us that the SEM is "rather crude" (p187) because it estimates an individual's observed score variation in the population (i.e., holding constant the true score). When examining the distribution of the SEM, examinees who score above the mean are more likely to have a positive error of measurement and examinees who score below the mean are more likely to have a negative error of measurement. Another point to consider is that, obviously, the further away from the mean that an individual scores on a given measure the larger the error of measurement (Hopkins, 1998). Finally, Thompson and Vacha-Haase (2000) pointed out that even if one chooses to use the SEM in an RG study it can only be useful when the same scale and form is used across studies because SEM is a function of the scale. In other words it would not make sense to look at the SEM if one was comparing studies that used different forms of a particular scale (forms with different variances) or if one was comparing multiple measures of the same construct.

Purpose of the Study

The purpose of this research was to examine the potential impact of selected methodological factors on the validity of conclusions from RG studies. Although all of the controversies described above are important, this study focused on four factors: (a) missing data in the primary studies, (b) transformation of the sample reliability estimates, (c) use of sample weights for estimating mean score reliability and building confidence bands, and (d) differences between analyses of score reliability estimates and estimates of SEM.

Method

The research was a Monte Carlo study in which random samples were simulated under known and controlled population conditions. In the Monte Carlo study, RG studies were simulated by generating samples in primary studies, estimating reliability of scores in these samples, and then aggregating the sample reliability estimates in the RG studies.

The Monte Carlo study included six factors in the design. These factors were (a) the true population reliability (with $\rho_{xx} = 0.40, 0.60, 0.80$ and 0.90), (b) sample size in the primary studies (with average sample sizes of 10, 50, 100, 500, and 1500), (c) number of primary studies in the RG study (with $k = 15, 50, 100$, and 150) (d) proportion of missing data (with proportions ranging from 0% to 90%), (e) homogeneity of the primary study samples (with score variance of $\sigma^2 = 1, 2, 4$, and 8) and (f) missing data mechanisms (randomly missing reliability estimates in the primary studies and systematically missing data such that the probability of missingness increased as the sample reliability estimate decreased).

Simulation of data. The research was conducted using SAS/IML version 8.1. Conditions for the study were run under Windows 98. Normally distributed random variables were generated using the RANNOR random number generator

in SAS. A different seed value for the random number generator was used in each execution of the program and the program code was verified by hand-checking results from benchmark datasets.

Measurement error was simulated in the data by generating two normally distributed random variables for each observation, one of which represents the 'true score' on the variable, the other representing measurement error. Fallible, observed scores on the variable were then calculated as the sum of the true and error components, consistent with classical measurement theory. The reliabilities of the scores were controlled by adjusting the error variance relative to the true score variance by:

$$\rho_{xx} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

where σ_T^2 and σ_E^2 are the true and error variances, respectively, and ρ_{xx} is the reliability.

For each simulated observation, two "observed scores" were generated by holding constant the random value for the true score component, but incorporating two, independent error score components. The two sets of observed scores provided a simulation of test-retest reliability estimation and the correlation between the two sets of scores provided the sample index of reliability. Similarly, the sample reliability index and the sample standard deviation were used to calculate a sample value of SEM.

For each condition investigated, 10,000 RG analyses were simulated. The use of 10,000 estimates provides adequate precision for the investigation of the bias in the reliability parameter estimates. For example, 10,000 samples provides a maximum 95% confidence interval width around an observed proportion that is ± .0098 (Robey & Barcikowski, 1992).

Conduct of RG analyses. Each RG analysis was conducted using the standard error of measurement (SEM) and the obtained sample reliability estimate. The latter was investigated in both its untransformed metric (i.e., r_{xx}), and using Fisher's z transformation:

$$z = \frac{1}{2}\ln\left(\frac{1+|r|}{1-|r|}\right) \text{ or } z = \tanh^{-1} r$$

to normalize the sampling distribution. Further, for RG studies based on reliability coefficients and those based on SEM, both unweighted analyses and weighted least squares analyses were conducted (Fuller & Hester, 1999; Raudenbush,1994; Hedges & Olkin, 1985).

Three treatments of missing data were applied to RG studies that included missing sample reliability estimates. In the listwise deletion approach, observations with missing reliability coefficients were deleted from the RG analysis. Such a listwise deletion approach is the strategy typically used with RG studies that have been conducted to date. In addition, two imputation procedures were applied to each simulated sample. In the simple regression imputation approach, an estimate of each missing reliability coefficient was obtained by first regressing (using cases with complete data) the observed reliability estimates on the observed sample variances. The obtained sample regression equation was then applied to the cases with missing reliability coefficients (i.e., using the sample variance for these cases) to obtain a predicted reliability coefficient. The predicted coefficients in the sample were then used in subsequent RG analyses. Finally, a multiple imputation approach was applied to each sample (Rubin, 1996; Schafer, 1997). The multiple imputation procedure replaces each missing value with a set of plausible values that represents the degree of uncertainty about the correct value to impute (that is, taking into account the uncertainty in the parameter estimates of the regression equation used to make imputations). This approach is an enhancement over simple imputation methods that fail to reflect the uncertainty about the predictions of the missing values, often resulting in point estimates of a variety of parameters that are not statistically valid. In general, multiple imputation inference involves three distinct phases (Schafer, 1997). First, missing data are filled in m times to generate complete data sets (for this research $m = 10$). Second, the m complete data sets are analyzed using standard statistical analyses (i.e., estimation of the mean and variance of the reliability estimates and the SEM). Finally, the results from the m complete data sets are combined to produce inferential results. That is, the mean value of the estimated mean reliability across the 10 imputations is used as the best estimate of the reliability coefficient in the population. In addition, the variance across the 10 imputations is used as an estimate of imputation variance. This additional source of variance is combined with the estimation variance of each imputation to produce a total variance which is used for the calculation of confidence intervals.

Evaluation of results. Each simulated RG study was used to obtain an estimated mean reliability and an estimated mean SEM. In addition, a 95% confidence band was constructed around each population estimate. For the construction of confidence bands, the sampling error of each estimate of score dependability index was calculated:

$$\hat{\sigma}_r^2 = \frac{1}{n-2}$$

$$\hat{\sigma}_z^2 = \frac{1}{\sqrt{n-3}}$$

$$\hat{\sigma}^2_{SEM} = \frac{1}{n-1}$$

where $\hat{\sigma}^2_r$, $\hat{\sigma}^2_z$ and $\hat{\sigma}^2_{SEM}$ are the estimated sampling variances of r_{xx}, Fisher's transformed r_{xx}, and SEM, respectively.

The standard error used for construction of the confidence band for the mean index of score dependability was obtained as

$$SE_\theta = \sqrt{\sum_{k=1}^{K}\left(\frac{1}{\sigma^2_{\theta k}}\right)^{-1}}$$

where $\sigma^2_{\theta k}$ is the sampling error variance for an index θ (i.e., r_{xx}, Fisher's transformed r_{xx}, or SEM) in the k^{th} study and the summation is across the studies included in the RG analysis.

The impact of the research design factors was evaluated based upon the bias in the mean estimates, the confidence band coverage, and the average confidence band width. Bias was estimated as the difference between the average sample estimate and the known population value of either the reliability coefficient or the SEM. That is,

$$Bias\left(\hat{\theta}\right) = \frac{\sum_{i}^{R}\left(\hat{\theta}_i - \theta\right)}{R}$$

where $\hat{\theta}_i$ = the sample estimate from the i^{th} RG study,

θ = the population value, and the summation is over the R simulated RG studies.

Confidence band coverage probabilities were estimated by computing the proportion of confidence bands in the R simulated RG studies that contained the parameter of interest. Similarly, confidence band width was computed as the average width of confidence bands from the R simulated RG studies.

Results

The results of this study were analyzed in terms of statistical bias of the estimates of $\rho_{xx'}$ and SEM, as well as the coverage probabilities of confidence bands and confidence band widths for the estimation of the reliability and SEM for the population as a whole.

Statistical Bias

Estimation of $\rho_{xx'}$. The distributions of bias estimates in regards to reliability across all conditions examined in this study are presented in Figure 1. For most conditions, the bias in the estimate of the mean reliability was relatively small (less than .05 in absolute value), but all methods suggest that specific conditions produced much larger biases in the estimate (on the order of .20). Further, most of the bias estimates were in a positive direction (the mean reliability was overestimated from the samples), with the exception of the multiple imputation missing data treatment applied to untransformed sample reliability values. For this method, both positive and negative biases were evident in the conditions examined. Further analyses of the bias in these estimates were approached by examining each of the design factors in this Monte Carlo study.

Estimates of the mean bias by the percentage and type of missing data are presented in Table 2. With no missing data, the mean bias did not exceed .01 for any of the approaches to reliability generalization (i.e., use of weighted estimators vs. non-weighted estimators, and use of Fisher's transformation vs. untransformed r values). Similar results of very low average bias were seen for randomly missing data and for systematically missing data, as long as no more the 30% of the observations presented missingness. As the proportion of randomly missing data increased, greater bias was evident in the multiple imputation approach to treating missing data (regardless of the use of weights or Fisher's transformation). For randomly missing data, the other approaches to missing data (regression imputation and listwise deletion) maintained unbiased estimates of the mean reliability. For systematically missing data, however, all of the methods of analysis evidenced increases in statistical bias of the estimate of the mean reliability.

The average estimates by the magnitude of $\rho_{xx'}$ are presented in Table 3. With no missing data, negligible bias was evident across all methods of analysis, regardless of the magnitude of the reliability in the population. With missing data present, however, all methods demonstrated increases in bias with lower values of $\rho_{xx'}$. With randomly missing data, only the multiple imputation method was affected by low values of $\rho_{xx'}$ (yielding average biases ranging from -.03 to .04), but with systematically missing data, all methods were affected (it should be noted, however, that the bias obtained with the multiple imputation approach was greater than those observed with other methods of treating missing data).

The estimates of bias by the number of studies included in the RG study (k) and the average sample size within these studies (n) are presented in Tables 4 – 6 for no missing data, randomly missing data, and systematically missing data, respectively. With no missing data (Table 4), all of the methods provided relatively unbiased estimates across values of k and n (with no biases exceeding .01 in absolute value). With randomly missing data (Table 5), the multiple imputation

missing data treatment showed the most bias across the majority of conditions and listwise deletion showed the least bias. The use of simple regression imputation was relatively unbiased except for conditions with small n and large k, in which the bias reached .02 in absolute value. Finally, with systematically missing data, all of the methods produced biased estimates of the mean reliability under conditions of small n and large k, with the bias being somewhat larger with analyses based on Fisher's transformation.

Estimation of SEM. The distributions of bias estimates in regards to SEM across all conditions examined in this study are presented in Figure 2. In contrast to the distribution of bias in the estimation of ρ_{xx}, the direction of the bias in these data depended upon the treatment of the missing data, with the MI method evidencing positive bias in some conditions, and the regression imputation and listwise deletion procedures evidencing a negative bias. However, for the majority of conditions examined, very little bias was present for any of the methods. Further analyses of the bias in these estimates were approached by examining each of the design factors in this Monte Carlo study.

Estimates of the mean bias in regards by the percentage and type of missing data are presented in Table 7. With no missing data, the average bias did not exceed .01. Similar results were seen for randomly missing data or systematically missing data, as long as no more the 30% of the observations presented missingness. As the proportion of randomly missing data increased, greater bias was evident in the multiple imputation approach to treating missing data and was as large as .11 for 90% missing data. For systematically missing data, the regression imputation and listwise deletion methods evidenced negative bias as the proportion of missing data increased.

The average estimates by the magnitude of ρ_{xx} are presented in Table 8. With no missing data the average bias remained close to zero (never exceeding .01 in absolute value). With missing data present, however, the bias increased as ρ_{xx} decreased, with the MI approach producing a positive bias with randomly missing data and the other two approaches evidencing negative bias with systematically missing data.

The estimates of bias by the number of studies included in the RG study (k) and the average sample size within these studies (n) are presented in Tables 9 – 11 for no missing data, randomly missing data, and systematically missing data, respectively. With no missing data (Table 9), all of the methods provided relatively unbiased estimates across values of k and n (with no biases exceeding .03 in absolute value). With randomly missing data (Table 10), the MI approach evidenced positive bias with small values of k, regardless of the size of n. Regression imputation and listwise deletion produced negative bias with small n conditions, but the bias was much smaller in magnitude. Finally, with systematically missing data (Table 11), all of the methods produced biased estimates with small n.

Confidence Band Coverage

Estimation of Population Mean Reliability. The distributions of estimated confidence band coverage across all conditions examined in this study are presented in Figure 3 in terms of untransformed reliability estimates and Fisher's transformations. For most conditions, the coverage obtained with analyses based on Fisher's transformation was superior to that obtained from the analysis of observed, untransformed reliability estimates. Further, the multiple imputation approach to missing data treatment provided better coverage probabilities than the listwise deletion or regression imputation approaches. Further analyses of the coverage probabilities were approached by examining each of the design factors in this Monte Carlo study.

The mean values of the estimated coverage probabilities by percentage and type of missing data are presented in Table 12. With no missing data, the Fisher bands provided conservative coverage (99%), while the bands based on untransformed r were excessively liberal (72% - 73%). With randomly missing data, the band coverage improved for listwise deletion, because this method produces wider confidence bands in the presence of missing data. However, the regression imputation approach showed notably worse coverage as the proportion of randomly missing data increased (reaching as low as 37% when the untransformed sample reliability coefficients were analyzed). With systematically missing data, the confidence bands obtained using listwise deletion and regression imputation with the sample reliability coefficients provided very poor coverage (less than 52%), and the use of Fisher's transformation with listwise deletion provided coverage as low as 78% with the largest proportions of missing data examined. In contrast, the use of multiple imputation with Fisher's transformation maintained band coverage at or above the nominal level, even with the most severe levels of systematically missing data.

Confidence band coverage probabilities by the magnitude of ρ_{xy} are provided in Table 13. The use of untransformed sample reliabilities provided notably poorer band coverage than the Fisher transformation. Further, the use of the multiple imputation approach provided substantially better confidence band coverage than the use of listwise deletion or regression imputation. With Fisher's transformation, the average band coverage with multiple imputation did not fall below 98%. In contrast, the use of listwise deletion provided band coverage as low as 90% with systematically missing data and Fisher's transformation, and as low as 48% without Fisher's transformation. Similarly, the use of regression imputation provided coverage as low as 68% with transformed reliabilities and as low as 40% with untransformed sample statistics.

The estimated confidence band coverage probabilities by k and n are presented in Table 14-16, for no missing data, randomly missing data, and systematically missing data, respectively. With no missing data (Table 14), the use of

Fisher's transformation provided conservative coverage probabilities (92% - 100%) across all of the conditions, while the untransformed sample reliabilities provided coverage closer to the nominal level as long as the sample sizes from the primary studies were large. With small samples, however, the confidence band coverage for untransformed sample reliability coefficients dropped to as low as low as 24% with the largest k and smallest n.

With randomly missing data (Table 15), the mean coverage for all analyses based on Fisher's transformation was superior to that of untransformed r, although the coverage was conservative in most conditions. For analyses based on the untransformed sample reliability coefficient, the use of multiple imputation provided bands with adequate coverage for all conditions except the largest k with the smallest n, in which the mean coverage dropped to only 92% - 93%. For the use of listwise deletion and regression imputation, coverage was poor with small sample sizes in the primary studies, becoming worse as k increased. However, adequate confidence band coverage with listwise deletion was obtained when the sample sizes in the primary studies were large.

For systematically missing data (Table 16), Fisher's transformation with multiple imputation provided conservative coverage, except under large k and small n conditions (dropping only as low as 88%). The other approaches evidenced much poorer band coverage, with the exception of large sample analyses, in which both listwise deletion and regression imputation provided adequate coverage if Fisher's transformation was used.

Estimation of Population Mean SEM. The distributions of estimated confidence band coverage for SEM across all conditions examined in this study are presented in Figure 4. For most conditions, the coverage obtained with analyses based on SEM was very poor, with analyses based on multiple imputations showing somewhat better coverage than those based on regression imputation or listwise deletion. Because confidence band coverage was so poor across the conditions, further analyses were not pursued.

Confidence Band Widths

Estimation of Population Mean Reliability. The distributions of estimated confidence band widths across all conditions examined in this study are presented in Figure 5.For most conditions, the widths of the bands based on multiple imputation were larger than those based on listwise deletion or regression imputation. Similarly, the bands obtained from the Fisher transformation were larger than those obtained from the untransformed sample reliability estimates, regardless of the missing data treatment applied. Further analyses of the band widths were approached by examining the design factors in this Monte Carlo study.

The mean values of the estimated confidence band widths by percentages and type of missing data are presented in Table 17. With no missing data, the average band width for Fisher's transformation (.06) was three times as large as that observed for the untransformed r (.02). With both randomly and systematically missing data the band widths increased with larger proportions of missing data for both listwise deletion and multiple imputations. With regression imputation, however, the average band width decreased with larger proportions of missing data.

Confidence band widths by the magnitude of $\rho_{xx'}$ are provided in Table 18. Across all conditions, the widths of the confidence bands decreased as the true reliability increased. Similarly, across all conditions, the widths of the bands constructed using Fisher's transformation were larger than those obtained from the untransformed sample reliability estimates.

Finally, the confidence band width by $\rho_{xx'}$ and n are presented in Tables 19 –21, for no missing data, randomly missing data and systematically missing data respectively. With no missing data (Table 19), the average confidence band widths were relatively small (never exceeding .19). Much wider confidence bands were evidenced by the multiple imputation procedure with randomly missing data (Table 20) and with systematically missing data (Table 21). Further, these confidence bands did not become appreciably smaller with larger values of *n*. The listwise deletion procedure provided slightly larger confidence bands with missing data, but the band widths decreased substantially with larger samples. Finally, the regression imputation procedure produced slightly smaller confidence bands in the presence of missing data.

Estimation of Population Mean SEM. The distributions of estimated confidence band widths across all conditions examined in this study are presented in Figure 6. As with the results for confidence bands constructed around the reliability estimates, the confidence bands for SEM were notably larger when the multiple imputation treatment of missing data was applied. The band width differences between listwise deletion and regression imputation were fairly small, although the bands for the former were somewhat wider on average. Further analyses of the band widths were approached by examining the design factors in this Monte Carlo study.

The mean values of the estimated mean confidence band widths by percentages and type of missing data are presented in Table 22. For both randomly and systematically missing data the band width increased dramatically for multiple imputations (0.26 - 1.27) as the proportion of missing data increased. Listwise deletion showed a much smaller increase in the average width of confidence bands (.03 - .07) and regression imputation evidenced a slight decrease in band widths.

Confidence band widths by the magnitude of ρ_{xx} are provided in Table 23. For all missing data conditions and all methods, the confidence bands decreased in width as the true reliability increased. However, the large differences in band width between the multiple imputation approach and the other approaches were evident across all values of ρ_{xx}.

Conclusions

In general, the results suggest that the use of Fisher's z transformation of the reliability estimates provided a modest increase in the accuracy of the estimation of the population mean reliability. Although the statistical bias in point estimates of the mean reliability were very small across most conditions, the confidence bands obtained using the Fisher transformation were more accurate than the confidence bands constructed using the untransformed r.

The importance of using the Fisher transformation in confidence band construction was especially evident with more challenging data conditions, such as RG studies based on a large number of primary studies but small samples in those studies, or RG studies conducted in the presence of missing data. Additionally, in these circumstances the use of weighted estimates provided slightly better confidence band coverage than the use of unweighted estimates. Finally, our comparison of missing data treatments suggests that the multiple imputation approach is superior to the listwise deletion approach, especially with the occurrence of systematically missing data. In fact, the results of this research suggest that the common practice in RG studies of listwise deletion for missing data can result in estimates that are extremely inaccurate. Although the confidence bands obtained using multiple imputation were substantially wider than those observed for listwise deletion, the increased width provided superior coverage probabilities. Further, the band width may be reduced by increasing the number of imputations that are used to obtain the estimates (this study employed 10 imputations for the missing data treatment).

Although the results for SEM suggest that unbiased estimates of the population mean may be obtained, the confidence band coverage was very poor across all conditions. This exceptionally poor coverage may result from the use of the product of two sample estimates that provide the sample SEM (i.e., the sample reliability coefficient and the sample standard deviation). Further research on the construction of confidence intervals for SEM is certainly needed.

In general, we consider RG studies as potentially being directed towards one of two goals: describing the distribution of score reliability indices in a corpus of published research reports, or estimating the properties of the distribution of score reliability in the hypothetical population of studies that have been and may be conducted using a given instrument. For simple descriptive applications, the use of sample weights or transformations is probably not needed. For

the estimation of population characteristics, however, methodological choices in the conduct of the RG study have a large impact on the accuracy of the inferences obtained.

While Thompson and Vacha-Haase (2000) believe that a series of RG studies could reveal that across samples, the reliability of scores for a given scale are relatively stable they also say that it is possible that such analyses could reveal that the variation in reliability is not related to research design factors. Regardless of the possible future uses and outcomes of the RG method, in order for these outcomes to have credibility the RG study design must have credibility. Thompson and Vacha-Haase (2000) defended their research design by stating:

> "We may not like the ingredients that go into making the RG sausage, but the RG chef can only work with the ingredients provided in the literature" (p 184)

In regards to this metaphor, the "RG chef" needs to also make sure that the ingredients are measured and prepared correctly so that the "RG sausage " doesn't leave the reader with an unpleasant aftertaste.

References

Capraro, M.M. Carpraro, R.M. & Henson. R.K (2001). Measurement error of scores on The Mathematics Anxiety Rating Scale across studies. *Educational and Psychological Measurement*, 61, 373- 386

Caruso, J. C. (2000). Reliability Generalization of NEO personality scales. *Education and Psychological Measurement*, 60, 236-254

Caruso, J.C. Witkiewitz, K, Belcourt-Diffloff, A., & Gottlieb, J (in press). Reliability of scores from the Eysenck Personality Questionnaire: A Reliability Generalization (RG) study. *Educational and Psychological Measurement,* 61

Crocker, L. & Angina, J (1986) *Introduction to classical and modern test theory.* Toronto: Holt. Rinehart & Winston.

Fuller, J. B., & Hester, K. (1999). Comparing the sample-weighted and unweighted meta-analysisi: An applied perspective. *Journal of Management, 25,* 803-828.

Glass, G.V. (1976). Primary, secondary, meta-analysis research. *Educational Researcher 5,* 3-8.

Hedges, L.V., & Olkin, I (1985) *Statistical Methods for Meta-Analysis.* San Diego, CA: Academic Press,Inc.

Helms, J.E. (1999) Another meta-analysis if the White Racial Identity Attitude Scale's Cronbach alphas: Implication for validity. *Measurement and Evaluation in Counseling and Development*, 32, 122-137.

Henson, R.K., Kogan, L.R., & Vacha-Hasse, T. (2001). A reliability generalization study of the Teacher Efficacy Scales and related instruments. *Educational and Psychological Measurement*, 61, p 404-420.

Henson, R. K., & Thompson, B.(2001, April). *Characterizing measurement error in test scores across studies: A tutorial on conducting "Reliability Generalization" Analysis.* Paper presented at the annual meeting of the American Educational Research Asociation, Seattle, WA

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Robey, R., & Barciowski, R. (1992) Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematics and Statistical Psychology 45* 283 –288.

Rosenthal, R (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638-641.

Rosenthal, R. (1994). Parametric measures of effect sizes. In H. Cooper. & L.V. Hedges (Eds.), *The Handbook of Research Synthesis.* (pp 231- 260). New York: Russell Sage Foundation.

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91, 473-489.

Sawilosky, S.S.(2000). Psychometrics versus datametrics: Comment on Vacha-Hasse's "reliability generalization" method and some ERM editorial policies. *Educational and Psychological Measurement, 60,* 157-173.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. New York: Chapman and Hall.

Schmidt F.L., & Hunter, J.E.(1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology,62,* 529-540.

Thompson, B., & Snyder, P..A. (1998). Statistical significance and reliability analyses in recent *JCD* research articles. *Journal of Counseling and Development, 76,* 436 –441.

Thompson, B. & Vacha-Hasse, T. (2000). Psychometrics *is* datametrics: The test is not reliable. . *Educational and Psychological Measurement,60,* 174- 195.

Vacha-Hasse, T., Kogan, L., Tani, C.R., & Woodall, R. A. (2001). Reliability generalization: Exploring reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement*, 61, 45-59

Vacha-Hasse, T., Kogan, L.R., & Thompson. B. (2000). Sample composition and variabilities in published studies versus the in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509-522.

Vacha- Hasse, T., Ness, C., Nilsson, J. & Rettz, D. (1999). Practices regarding reporting of reliability coefficients: A review if three journals. *Journal of Experimental Education*, 67, 335-341.

Vacha-Hasse, T., Tani, C.R., Kogan, L.R., Woodall, R.A. & Thompson, B. (in press) Reliability generalization: Exploring Reliability variations on MMPI/MMPI-2 Validity scale scores. *Assessment.*

Vacha-Hasse, Y (1998). Reliability generalization: exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, 58, 6- 20.

Viswesvaran, C., & Ones, D. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability Generalization across studies and measures. *Educational and Psychological Measurement*, *60*, 24-235.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability Generalization across studies. *Educational and Psychological Measurement*, 61,201 –223.

Table 1

Current List of Published Reliability Generalization Studies.

Capraro, M.M. Carpraro, R.M. & Henson. R.K (2001). Measurement error of scores on The Mathematics Anxiety Rating Scale across studies. *Educational and Psychological Measurement,61*, 373- 386
Caruso, J. C. (2000). Reliability Generalization of NEO personality scales. *Education and Psychological Measurement, 60*, 236-254
Caruso, J.C. Witkiewitz, K, Belcourt-Diffloff, A., & Gottlieb, J (in press). Reliability of scores from the Eysenck Personality Questionnaire: A Reliability Generalization (RG) study. *Educational and Psychological Measurement, 61*
Helms, J.E. (1999) Another meta-analysis if the White Racial Identity Attitude Scale's Cronbach alphas: Implication for validity. *Measurement and Evaluation in Counseling and Development, 32, 122-137.*
Henson, R.K., Kogan, L.R., & Vacha-Hasse, T. (2001). A reliability generalization study of the Teacher Efficacy Scales and related instruments. *Educational and Psychological Measurement, 61*, p 404-420.
Vacha-Hasse, T., Kogan, L., Tani, C.R., & Woodall, R. A. (2001). Reliability generalization: Exploring reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement, 61, 45-59*
Vacha-Hasse, T., Tani, C.R., Kogan, L.R., Woodall, R.A. & Thompson, B. (in press) Reliability generalization: Exploring Reliability variations on MMPI/MMPI-2 Validity scale scores. *Assessment.*
Vacha-Hasse, Y (1998). Reliability generalization: exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6- 20.
Viswesvaran, C., & Ones, D. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability Generalization across studies and measures. *Educational and Psychological Measurement, 60,* 24-235.
Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability Generalization across studies. *Educational and Psychological Measurement, 61,*201 –223.

Table 2
Bias in Estimated Mean Reliability by Percentage and Type of Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
Missingness	Percent	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0.0	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01
Random	0.3	0.00	0.00	0.00	0.00	0.01	0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
	0.6	0.01	0.01	0.00	0.00	0.02	0.02	-0.01	-0.01	0.00	0.00	0.00	0.00
	0.9	0.01	0.01	-0.01	-0.01	0.04	0.04	-0.03	-0.03	0.00	0.00	0.00	0.00
Systematic	0.3	0.01	0.01	0.00	0.00	0.01	0.01	-0.01	-0.01	0.01	0.01	0.00	0.00
	0.6	0.02	0.02	0.01	0.01	0.03	0.03	0.00	0.00	0.02	0.02	0.01	0.01
	0.9	0.05	0.05	0.05	0.05	0.08	0.08	0.03	0.03	0.06	0.06	0.05	0.05

Table 3
Bias in Estimated Mean Reliability by Population Mean Reliability and Type of Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
Missingness	Rhoxx	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0.4	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01
	0.6	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01
	0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Random	0.4	0.01	0.01	-0.01	-0.01	0.04	0.04	-0.03	-0.03	0.01	0.01	-0.01	-0.01
	0.6	0.01	0.01	-0.01	-0.01	0.03	0.03	-0.02	-0.02	0.00	0.00	0.00	-0.01
	0.8	0.00	0.00	0.00	0.00	0.01	0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
	0.9	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
Systematic	0.4	0.04	0.04	0.03	0.03	0.07	0.07	0.02	0.02	0.04	0.04	0.03	0.03
	0.6	0.03	0.03	0.02	0.02	0.05	0.05	0.01	0.01	0.03	0.03	0.02	0.02
	0.8	0.01	0.01	0.01	0.01	0.02	0.02	0.00	0.00	0.02	0.02	0.01	0.01
	0.9	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.01

Table 4
Bias in Estimated Mean Reliability by Number of Studies and Sample Size for No Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
K	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
15	10	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01
	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50	10	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01
	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100	10	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01
	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
150	10	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01
	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5

Bias in Estimated Mean Reliability by Number of Studies and Sample Size for Randomly Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
K	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
15	10	0.02	0.02	-0.02	-0.02	0.02	0.02	-0.04	-0.04	0.01	0.01	-0.01	-0.01
	50	0.00	0.00	0.00	0.00	0.02	0.02	-0.02	-0.02	0.00	0.00	0.00	0.00
	100	0.00	0.00	0.00	0.00	0.02	0.02	-0.02	-0.02	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.02	0.02	-0.02	-0.02	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.02	0.02	-0.02	-0.02	0.00	0.00	0.00	0.00
50	10	0.02	0.02	-0.02	-0.02	0.03	0.03	-0.04	-0.04	0.01	0.01	-0.01	-0.01
	50	0.00	0.00	0.00	0.00	0.02	0.02	-0.02	-0.02	0.00	0.00	0.00	0.00
	100	0.00	0.00	0.00	0.00	0.02	0.02	-0.02	-0.02	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.02	0.02	-0.02	-0.02	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.02	0.02	-0.02	-0.02	0.00	0.00	0.00	0.00
100	10	0.02	0.02	-0.02	-0.02	0.03	0.03	-0.03	-0.03	0.01	0.01	-0.01	-0.01
	50	0.00	0.00	0.00	0.00	0.02	0.02	-0.01	-0.01	0.00	0.00	0.00	0.00
	100	0.00	0.00	0.00	0.00	0.02	0.02	-0.01	-0.01	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.02	0.02	-0.01	-0.01	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.02	0.02	-0.01	-0.01	0.00	0.00	0.00	0.00
150	10	0.02	0.02	-0.02	-0.02	0.03	0.03	-0.02	-0.02	0.01	0.01	-0.01	-0.01
	50	0.00	0.00	0.00	0.00	0.02	0.02	-0.01	-0.01	0.00	0.00	0.00	0.00
	100	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00

Table 6
Bias in Estimated Mean Reliability by Number of Studies and Sample Size for Systematically Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
K	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
15	10	0.03	0.03	0.00	0.00	0.04	0.04	-0.02	-0.02	0.03	0.03	0.01	0.01
	50	0.01	0.01	0.00	0.00	0.03	0.03	-0.01	-0.01	0.01	0.01	0.01	0.01
	100	0.01	0.01	0.00	0.00	0.03	0.03	-0.01	-0.01	0.01	0.01	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.03	0.03	-0.01	-0.01	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.03	0.03	-0.01	-0.01	0.00	0.00	0.00	0.00
50	10	0.06	0.06	0.03	0.03	0.06	0.06	0.02	0.02	0.06	0.06	0.04	0.04
	50	0.02	0.02	0.02	0.02	0.04	0.04	0.01	0.01	0.02	0.02	0.02	0.02
	100	0.01	0.01	0.01	0.01	0.03	0.03	0.00	0.00	0.02	0.02	0.01	0.01
	500	0.01	0.01	0.01	0.01	0.03	0.03	-0.01	-0.01	0.01	0.01	0.01	0.01
	1500	0.00	0.00	0.00	0.00	0.02	0.02	-0.01	-0.01	0.00	0.00	0.00	0.00
100	10	0.06	0.06	0.03	0.03	0.06	0.06	0.03	0.03	0.06	0.06	0.04	0.04
	50	0.02	0.02	0.02	0.02	0.04	0.04	0.01	0.01	0.02	0.02	0.02	0.02
	100	0.02	0.02	0.01	0.01	0.03	0.03	0.01	0.01	0.02	0.02	0.02	0.02
	500	0.01	0.01	0.01	0.01	0.02	0.02	0.00	0.00	0.01	0.01	0.01	0.01
	1500	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
150	10	0.06	0.06	0.03	0.03	0.06	0.06	0.03	0.03	0.06	0.06	0.04	0.04
	50	0.02	0.02	0.02	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02
	100	0.02	0.02	0.01	0.01	0.03	0.03	0.01	0.01	0.02	0.02	0.02	0.02
	500	0.01	0.01	0.01	0.01	0.02	0.02	0.00	0.00	0.01	0.01	0.01	0.01
	1500	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00

Table 7
Bias in Estimated Mean Standard Error of Measurement by Percentage and Type of Missing Data.

		Regression Imputation		Multiple Imputation		Listwise Deletion	
		SEM	SEM	SEM	SEM	SEM	SEM
Missingness	Percent	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Random	0.3	-0.01	-0.01	0.00	0.00	-0.01	-0.01
	0.6	-0.01	-0.01	0.03	0.03	-0.01	-0.01
	0.9	-0.01	-0.01	0.11	0.11	-0.01	-0.01
Systematic	0.3	-0.01	-0.01	0.00	0.00	-0.01	-0.01
	0.6	-0.03	-0.03	0.01	0.01	-0.03	-0.02
	0.9	-0.09	-0.09	0.00	0.00	-0.07	-0.07

Table 8
Bias in Estimated Mean Standard Error of Measurement by Population Mean Reliability and Type of Missing Data.

		Regression Imputation		Multiple Imputation		Listwise Deletion	
		SEM	SEM	SEM	SEM	SEM	SEM
Missingness	Rhoxx	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0.4	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	0.6	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	0.8	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	0.9	0.00	0.00	0.00	0.00	0.00	0.00
Random	0.4	-0.01	-0.01	0.08	0.08	-0.01	-0.01
	0.6	-0.01	-0.01	0.05	0.05	-0.01	-0.01
	0.8	-0.01	-0.01	0.03	0.03	-0.01	-0.01
	0.9	0.00	0.00	0.02	0.02	0.00	0.00
Systematic	0.4	-0.06	-0.06	0.01	0.01	-0.05	-0.05
	0.6	-0.04	-0.04	0.01	0.01	-0.04	-0.04
	0.8	-0.03	-0.03	0.00	0.00	-0.02	-0.02
	0.9	-0.02	-0.02	0.00	0.00	-0.02	-0.02

Table 9
Bias in Estimated Mean Standard Error of Measurement by Number of Studies and Sample Size for No Missing Data.

		Regression Imputation		Multiple Imputation		Listwise Deletion	
		SEM	SEM	SEM	SEM	SEM	SEM
K	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
15	10	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
	50	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	100	0.00	0.00	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.00	0.00
50	10	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
	50	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	100	0.00	0.00	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.00	0.00
100	10	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
	50	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	100	0.00	0.00	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.00	0.00
150	10	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
	50	-0.01	0.00	-0.01	0.00	-0.01	0.00
	100	0.00	0.00	0.00	0.00	0.00	0.00
	500	0.00	0.00	0.00	0.00	0.00	0.00
	1500	0.00	0.00	0.00	0.00	0.00	0.00

Table 10
Bias in Estimated Mean Standard Error of Measurement by Number of Studies and Sample Size for Randomly Missing Data.

		Regression Imputation		Multiple Imputation		Listwise Deletion	
		SEM	SEM	SEM	SEM	SEM	SEM
K	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
15	10	-0.03	-0.03	0.05	0.06	-0.03	-0.03
	50	0.00	0.00	0.07	0.07	-0.01	-0.01
	100	0.00	0.00	0.07	0.07	0.00	0.00
	500	0.00	0.00	0.08	0.08	0.00	0.00
	1500	0.00	0.00	0.08	0.08	0.00	0.00
50	10	-0.02	-0.02	0.06	0.07	-0.03	-0.03
	50	0.00	0.00	0.08	0.08	-0.01	-0.01
	100	0.00	0.00	0.08	0.08	0.00	0.00
	500	0.00	0.00	0.08	0.08	0.00	0.00
	1500	0.00	0.00	0.08	0.08	0.00	0.00
100	10	-0.03	-0.03	0.01	0.01	-0.03	-0.03
	50	0.00	0.00	0.03	0.03	-0.01	-0.01
	100	0.00	0.00	0.03	0.03	0.00	0.00
	500	0.00	0.00	0.03	0.03	0.00	0.00
	1500	0.00	0.00	0.03	0.03	0.00	0.00
150	10	-0.03	-0.03	-0.01	-0.01	-0.03	-0.03
	50	-0.01	-0.01	0.01	0.02	-0.01	-0.01
	100	0.00	0.00	0.02	0.02	0.00	0.00
	500	0.00	0.00	0.02	0.02	0.00	0.00
	1500	0.00	0.00	0.02	0.02	0.00	0.00

Table 11
Bias in Estimated Mean Standard Error of Measurement by Number of Studies and Sample Size for Systematically Missing Data.

		Regression Imputation		Multiple Imputation		Listwise Deletion	
		SEM	SEM	SEM	SEM	SEM	SEM
K	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
15	10	-0.06	-0.05	0.02	0.03	-0.06	-0.05
	50	-0.02	-0.02	0.06	0.06	-0.02	-0.02
	100	-0.01	-0.01	0.06	0.06	-0.01	-0.01
	500	0.00	0.00	0.07	0.07	0.00	0.00
	1500	0.00	0.00	0.07	0.07	0.00	0.00
50	10	-0.11	-0.11	-0.04	-0.04	-0.10	-0.10
	50	-0.04	-0.04	0.02	0.02	-0.03	-0.03
	100	-0.03	-0.03	0.04	0.04	-0.02	-0.02
	500	-0.01	-0.01	0.05	0.05	-0.01	-0.01
	1500	-0.01	-0.01	0.06	0.06	-0.01	-0.01
100	10	-0.12	-0.12	-0.09	-0.09	-0.10	-0.10
	50	-0.04	-0.04	-0.02	-0.02	-0.04	-0.03
	100	-0.03	-0.03	-0.01	-0.01	-0.02	-0.02
	500	-0.01	-0.01	0.01	0.01	-0.01	-0.01
	1500	-0.01	-0.01	0.02	0.02	-0.01	-0.01
150	10	-0.12	-0.12	-0.10	-0.10	-0.10	-0.10
	50	-0.04	-0.04	-0.03	-0.03	-0.04	-0.03
	100	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02
	500	-0.01	-0.01	0.00	0.00	-0.01	-0.01
	1500	-0.01	-0.01	0.01	0.01	-0.01	-0.01

Table 12
Confidence Band Coverage for Mean Reliability by Percentage and Type of Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
Missingness	Percent	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0	0.99	0.99	0.79	0.78	0.99	0.99	0.73	0.72	0.99	0.99	0.73	0.72
Random	0.3	0.98	0.98	0.7	0.7	1	1	0.97	0.97	1	1	0.81	0.8
	0.6	0.95	0.95	0.61	0.6	1	1	1	1	1	1	0.85	0.83
	0.9	0.86	0.86	0.37	0.37	1	1	1	1	1	1	0.88	0.87
Systematic	0.3	0.97	0.97	0.72	0.72	1	1	0.98	0.97	0.99	0.99	0.84	0.83
	0.6	0.84	0.85	0.42	0.41	1	1	0.99	0.99	0.93	0.92	0.52	0.5
	0.9	0.17	0.17	0.01	0.01	0.95	0.95	0.89	0.89	0.79	0.78	0.03	0.03

Table 13

Confidence Band Coverage for Mean Reliability by Population Mean Reliability and Type of Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
Missingness	Rhoxx	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0.4	0.99	1	0.84	0.85	0.99	0.99	0.78	0.8	0.99	0.99	0.78	0.8
	0.6	0.99	0.99	0.78	0.8	0.99	0.99	0.75	0.76	0.99	0.99	0.75	0.76
	0.8	0.99	0.99	0.76	0.77	0.99	0.99	0.71	0.72	0.99	0.99	0.71	0.72
	0.9	1	1	0.72	0.73	0.99	0.99	0.67	0.68	0.99	0.99	0.67	0.68
Random	0.4	0.92	0.92	0.59	0.59	1	1	1	1	1	1	0.87	0.88
	0.6	0.92	0.92	0.57	0.58	1	1	0.99	0.99	1	1	0.84	0.85
	0.8	0.94	0.94	0.57	0.57	1	1	0.99	0.99	1	1	0.82	0.83
	0.9	0.96	0.96	0.56	0.56	1	1	0.98	0.98	1	1	0.8	0.81
Systematic	0.4	0.7	0.7	0.4	0.4	0.99	0.99	0.97	0.97	0.91	0.92	0.5	0.51
	0.6	0.68	0.68	0.41	0.41	0.99	0.99	0.96	0.97	0.9	0.9	0.49	0.5
	0.8	0.7	0.71	0.42	0.42	0.98	0.98	0.96	0.96	0.9	0.9	0.48	0.49
	0.9	0.74	0.74	0.42	0.43	0.98	0.98	0.94	0.94	0.93	0.93	0.48	0.49

Table 14
Confidence Band Coverage for Mean Reliability by Number of Studies and Sample Size for No Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
K	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
15	10	1.00	1.00	0.64	0.66	1.00	1.00	0.64	0.66	1.00	1.00	0.64	0.66
	50	1.00	1.00	0.92	0.92	1.00	1.00	0.92	0.92	1.00	1.00	0.92	0.92
	100	1.00	1.00	0.93	0.94	1.00	1.00	0.93	0.94	1.00	1.00	0.93	0.94
	500	1.00	1.00	0.94	0.95	1.00	1.00	0.94	0.95	1.00	1.00	0.94	0.95
	1500	1.00	1.00	0.95	0.96	1.00	1.00	0.95	0.96	1.00	1.00	0.95	0.96
50	10	0.99	0.99	0.46	0.48	0.99	0.99	0.46	0.48	0.99	0.99	0.46	0.48
	50	1.00	1.00	0.89	0.9	1.00	1.00	0.89	0.9	1.00	1.00	0.89	0.9
	100	1.00	1.00	0.92	0.93	1.00	1.00	0.92	0.93	1.00	1.00	0.92	0.93
	500	1.00	1.00	0.94	0.95	1.00	1.00	0.94	0.95	1.00	1.00	0.94	0.95
	1500	1.00	1.00	0.95	0.96	1.00	1.00	0.95	0.96	1.00	1.00	0.95	0.96
100	10	0.96	0.97	0.33	0.35	0.96	0.97	0.33	0.35	0.96	0.97	0.33	0.35
	50	1.00	1.00	0.85	0.87	1.00	1.00	0.85	0.87	1.00	1.00	0.85	0.87
	100	1.00	1.00	0.9	0.91	1.00	1.00	0.91	0.92	1.00	1.00	0.91	0.92
	500	1.00	1.00	0.94	0.95	1.00	1.00	0.94	0.95	1.00	1.00	0.94	0.95
	1500	1.00	1.00	0.95	0.96	1.00	1.00	0.95	0.96	1.00	1.00	0.95	0.96
150	10	0.92	0.94	0.24	0.27	0.92	0.94	0.24	0.27	0.92	0.94	0.24	0.27
	50	1.00	1.00	0.82	0.83	1.00	1.00	0.84	0.85	1.00	1.00	0.83	0.84
	100	1.00	1.00	0.89	0.90	1.00	1.00	0.9	0.91	1.00	1.00	0.89	0.90
	500	1.00	1.00	0.93	0.94	1.00	1.00	0.94	0.94	1.00	1.00	0.94	0.94
	1500	1.00	1.00	0.95	0.95	1.00	1.00	0.95	0.95	1.00	1.00	0.95	0.95

Table 15
Confidence Band Coverage for Mean Reliability by Number of Studies and Sample Size for Randomly Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
K	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
15	10	0.94	0.94	0.47	0.48	1.00	1.00	0.97	0.98	1.00	1.00	0.71	0.73
	50	1.00	1.00	0.75	0.76	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.93
	100	1.00	1.00	0.79	0.80	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.94
	500	1.00	1.00	0.81	0.82	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.95
	1500	1.00	1.00	0.82	0.82	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.95
50	10	0.81	0.82	0.25	0.25	1.00	1.00	0.96	0.97	1.00	1.00	0.63	0.64
	50	0.95	0.95	0.63	0.63	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.92
	100	0.98	0.98	0.68	0.68	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.94
	500	1.00	1.00	0.71	0.71	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.95
	1500	1.00	1.00	0.71	0.72	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.95
100	10	0.76	0.76	0.16	0.16	1.00	1.00	0.94	0.95	0.99	0.99	0.53	0.55
	50	0.96	0.96	0.61	0.61	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.90
	100	0.98	0.98	0.68	0.69	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.93
	500	1.00	1.00	0.72	0.72	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.95
	1500	1.00	1.00	0.72	0.73	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.95
150	10	0.69	0.7	0.11	0.12	1.00	1.00	0.92	0.93	0.98	0.98	0.47	0.49
	50	0.96	0.96	0.59	0.6	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.89
	100	0.99	0.99	0.68	0.69	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.93
	500	1.00	1.00	0.72	0.73	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.95
	1500	1.00	1.00	0.73	0.73	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.95

Table 16
Confidence Band Coverage for Mean Reliability by Number of Studies and Sample Size for Systematically Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
K	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
15	10	0.89	0.9	0.45	0.46	1.00	1.00	0.96	0.96	0.99	1.00	0.66	0.67
	50	0.99	0.99	0.72	0.73	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.88
	100	1.00	1.00	0.76	0.76	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.9
	500	1.00	1.00	0.78	0.78	1.00	1.00	1.00	1.00	1.00	1.00	0.9	0.91
	1500	1.00	1.00	0.79	0.79	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.92
50	10	0.53	0.53	0.2	0.21	0.95	0.96	0.88	0.88	0.84	0.86	0.35	0.36
	50	0.71	0.72	0.45	0.46	1.00	1.00	0.99	0.99	1.00	1.00	0.54	0.55
	100	0.77	0.77	0.48	0.48	1.00	1.00	1.00	1.00	1.00	1.00	0.56	0.57
	500	0.92	0.92	0.49	0.49	1.00	1.00	1.00	1.00	1.00	1.00	0.57	0.58
	1500	0.98	0.98	0.49	0.50	1.00	1.00	1.00	1.00	1.00	1.00	0.57	0.58
100	10	0.39	0.4	0.14	0.14	0.92	0.92	0.81	0.82	0.55	0.57	0.26	0.28
	50	0.63	0.63	0.4	0.4	1.00	1.00	1.00	1.00	0.97	0.98	0.44	0.45
	100	0.67	0.68	0.42	0.42	1.00	1.00	1.00	1.00	1.00	1.00	0.44	0.45
	500	0.78	0.78	0.43	0.43	1.00	1.00	1.00	1.00	1.00	1.00	0.44	0.45
	1500	0.92	0.92	0.42	0.43	1.00	1.00	1.00	1.00	1.00	1.00	0.43	0.45
150	10	0.32	0.33	0.1.00	0.11	0.88	0.89	0.74	0.75	0.43	0.44	0.21	0.23
	50	0.59	0.59	0.37	0.37	1.00	1.00	0.99	0.99	0.84	0.86	0.39	0.40
	100	0.65	0.65	0.39	0.39	1.00	1.00	1.00	1.00	0.98	0.99	0.39	0.40
	500	0.7	0.7	0.39	0.39	1.00	1.00	1.00	1.00	1.00	1.00	0.38	0.39
	1500	0.81	0.81	0.38	0.39	1.00	1.00	1.00	1.00	1.00	1.00	0.38	0.39

Table 17
Confidence Band Width for Mean Relibility by Percentage and Type of Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
Missingness	Percent	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0	0.06	0.06	0.02	0.02	0.06	0.06	0.02	0.02	0.06	0.06	0.02	0.02
Random	30	0.06	0.06	0.02	0.02	0.16	0.16	0.12	0.12	0.07	0.07	0.02	0.02
	60	0.06	0.06	0.02	0.02	0.36	0.36	0.28	0.28	0.09	0.09	0.03	0.03
	90	0.04	0.04	0.01	0.01	0.71	0.71	0.6	0.6	0.14	0.14	0.05	0.05
Systematic	30	0.06	0.06	0.02	0.02	0.16	0.16	0.12	0.12	0.07	0.07	0.02	0.02
	60	0.05	0.05	0.01	0.01	0.33	0.33	0.26	0.26	0.08	0.08	0.03	0.03
	90	0.03	0.03	0.01	0.01	0.51	0.51	0.40	0.40	0.10	0.10	0.03	0.03

Table 18
Confidence Band Width for Mean Reliability by Population Mean Reliability.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
Missingness	Rhoxx	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0.4	0.11	0.11	0.03	0.03	0.11	0.11	0.03	0.03	0.11	0.11	0.03	0.03
	0.6	0.07	0.07	0.02	0.02	0.07	0.07	0.02	0.02	0.07	0.07	0.02	0.02
	0.8	0.04	0.04	0.01	0.01	0.04	0.04	0.01	0.01	0.04	0.04	0.01	0.01
	0.9	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01
Random	0.4	0.10	0.10	0.03	0.03	0.71	0.71	0.60	0.60	0.18	0.18	0.06	0.06
	0.6	0.06	0.06	0.02	0.02	0.48	0.48	0.38	0.38	0.12	0.12	0.04	0.04
	0.8	0.03	0.03	0.01	0.01	0.23	0.23	0.18	0.18	0.06	0.06	0.02	0.02
	0.9	0.02	0.02	0.01	0.01	0.11	0.11	0.08	0.08	0.03	0.03	0.01	0.01
Systematic	0.4	0.09	0.09	0.03	0.03	0.60	0.60	0.47	0.47	0.16	0.16	0.05	0.05
	0.6	0.06	0.06	0.02	0.02	0.39	0.39	0.30	0.30	0.10	0.10	0.03	0.03
	0.8	0.03	0.03	0.01	0.01	0.19	0.19	0.14	0.14	0.05	0.05	0.01	0.01
	0.9	0.02	0.02	0.01	0.01	0.08	0.08	0.07	0.07	0.02	0.02	0.01	0.01

Table 19
Confidence Band Width by Population Mean Reliability and Sample Size for No Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
Rhox	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
0.4	10	0.19	0.19	0.08	0.08	0.19	0.19	0.08	0.08	0.19	0.19	0.08	0.08
	50	0.12	0.12	0.04	0.04	0.12	0.12	0.04	0.04	0.12	0.12	0.04	0.04
	100	0.10	0.10	0.03	0.03	0.10	0.10	0.03	0.03	0.10	0.10	0.03	0.03
	500	0.07	0.07	0.02	0.02	0.07	0.07	0.02	0.02	0.07	0.07	0.02	0.02
	1500	0.05	0.05	0.01	0.01	0.05	0.05	0.01	0.01	0.05	0.05	0.01	0.01
0.6	10	0.12	0.12	0.04	0.04	0.12	0.12	0.04	0.04	0.12	0.12	0.04	0.04
	50	0.09	0.09	0.03	0.03	0.09	0.09	0.03	0.03	0.09	0.09	0.03	0.03
	100	0.07	0.07	0.02	0.02	0.07	0.07	0.02	0.02	0.07	0.07	0.02	0.02
	500	0.05	0.05	0.01	0.01	0.05	0.05	0.01	0.01	0.05	0.05	0.01	0.01
	1500	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01
0.8	10	0.06	0.06	0.02	0.02	0.06	0.06	0.02	0.02	0.06	0.06	0.02	0.02
	50	0.04	0.04	0.02	0.02	0.05	0.05	0.02	0.02	0.04	0.04	0.02	0.02
	100	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01
	500	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01
	1500	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01
0.9	10	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01
	50	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01
	100	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.01
	500	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	1500	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 20
Confidence Band Width for Mean Reliability by Population Mean Reliability and Sample Size for Randomly Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
Rhox	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
0.4	10	0.17	0.17	0.05	0.05	0.74	0.74	0.62	0.62	0.32	0.32	0.13	0.13
	50	0.11	0.11	0.04	0.04	0.71	0.71	0.60	0.60	0.20	0.20	0.08	0.08
	100	0.10	0.10	0.03	0.03	0.71	0.71	0.59	0.59	0.17	0.17	0.05	0.05
	500	0.07	0.07	0.01	0.01	0.70	0.70	0.59	0.59	0.11	0.11	0.02	0.02
	1500	0.05	0.05	0.01	0.01	0.70	0.70	0.59	0.59	0.09	0.09	0.01	0.01
0.6	10	0.11	0.11	0.03	0.03	0.49	0.49	0.40	0.40	0.21	0.21	0.08	0.08
	50	0.08	0.08	0.03	0.03	0.48	0.48	0.38	0.38	0.13	0.13	0.05	0.05
	100	0.07	0.07	0.02	0.02	0.48	0.48	0.37	0.37	0.11	0.11	0.03	0.03
	500	0.04	0.04	0.01	0.01	0.48	0.48	0.37	0.37	0.08	0.08	0.02	0.02
	1500	0.03	0.03	0.01	0.01	0.48	0.48	0.37	0.37	0.06	0.06	0.01	0.01
0.8	10	0.06	0.06	0.01	0.01	0.24	0.24	0.19	0.2	0.11	0.11	0.04	0.04
	50	0.04	0.04	0.01	0.01	0.23	0.23	0.17	0.17	0.07	0.07	0.02	0.02
	100	0.03	0.03	0.01	0.01	0.24	0.24	0.17	0.17	0.06	0.06	0.02	0.02
	500	0.02	0.02	0.01	0.01	0.23	0.23	0.17	0.17	0.04	0.04	0.01	0.01
	1500	0.02	0.02	0.01	0.01	0.23	0.23	0.17	0.17	0.03	0.03	0.01	0.01
0.9	10	0.03	0.03	0.01	0.01	0.11	0.11	0.09	0.09	0.05	0.05	0.02	0.02
	50	0.02	0.02	0.01	0.01	0.11	0.11	0.08	0.08	0.03	0.04	0.01	0.01
	100	0.02	0.02	0.01	0.01	0.10	0.11	0.08	0.08	0.03	0.03	0.01	0.01
	500	0.01	0.01	0.01	0.01	0.10	0.10	0.08	0.08	0.02	0.02	0.01	0.01
	1500	0.01	0.01	0.01	0.01	0.10	0.10	0.08	0.08	0.02	0.02	0.01	0.01

Table 21
Confidence Band Width for Mean Reliability by Population Mean Reliability and Sample Size for Systematically Missing Data.

		Regression Imputation				Multiple Imputation				Listwise Deletion			
		Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r	Fisher Z	Fisher Z	Sample r	Sample r
Rhox	N	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
0.4	10	0.15	0.15	0.05	0.05	0.53	0.54	0.4	0.41	0.24	0.25	0.11	0.11
	50	0.11	0.11	0.04	0.04	0.6	0.6	0.47	0.47	0.19	0.19	0.07	0.07
	100	0.09	0.09	0.03	0.03	0.61	0.61	0.48	0.48	0.16	0.16	0.05	0.05
	500	0.06	0.06	0.01	0.01	0.62	0.62	0.50	0.50	0.11	0.11	0.02	0.02
	1500	0.05	0.05	0.01	0.01	0.63	0.63	0.51	0.51	0.09	0.09	0.01	0.01
0.6	10	0.09	0.09	0.03	0.03	0.33	0.33	0.25	0.25	0.15	0.16	0.06	0.06
	50	0.07	0.07	0.03	0.03	0.39	0.39	0.29	0.29	0.12	0.12	0.04	0.04
	100	0.06	0.06	0.02	0.02	0.4	0.4	0.3	0.3	0.10	0.10	0.03	0.03
	500	0.04	0.04	0.01	0.01	0.42	0.42	0.32	0.32	0.07	0.07	0.02	0.02
	1500	0.03	0.03	0.01	0.01	0.43	0.43	0.33	0.33	0.06	0.06	0.01	0.01
0.8	10	0.05	0.05	0.01	0.01	0.14	0.15	0.12	0.12	0.08	0.08	0.03	0.03
	50	0.04	0.04	0.01	0.01	0.18	0.18	0.14	0.14	0.06	0.06	0.02	0.02
	100	0.03	0.03	0.01	0.01	0.19	0.19	0.14	0.14	0.05	0.05	0.02	0.02
	500	0.02	0.02	0.01	0.01	0.2	0.2	0.15	0.15	0.04	0.04	0.01	0.01
	1500	0.02	0.02	0.01	0.01	0.21	0.21	0.15	0.15	0.03	0.03	0.01	0.01
0.9	10	0.03	0.03	0.01	0.01	0.06	0.06	0.06	0.06	0.04	0.04	0.01	0.01
	50	0.02	0.02	0.01	0.01	0.08	0.08	0.07	0.07	0.03	0.03	0.01	0.01
	100	0.02	0.02	0.01	0.01	0.09	0.09	0.07	0.07	0.02	0.02	0.01	0.01
	500	0.01	0.01	0.01	0.01	0.09	0.09	0.07	0.07	0.02	0.02	0.01	0.01
	1500	0.01	0.01	0.01	0.01	0.09	0.09	0.07	0.07	0.01	0.01	0.01	0.01

Table 22
Confidence Band Width for Standard Error of Measurement by Percentage and Type of Missing Data.

		Regression Imputation		Multiple Imputation		Listwise Deletion	
		SEM	SEM	SEM	SEM	SEM	SEM
Missingness	Percent	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0	0.06	0.03	0.03	0.03	0.03	0.03
Random	0.3	0.06	0.03	0.26	0.26	0.03	0.03
	0.6	0.06	0.03	0.65	0.65	0.05	0.05
	0.9	0.04	0.02	1.27	1.27	0.07	0.07
Systematic							
	0.3	0.06	0.03	0.26	0.26	0.03	0.03
	0.6	0.05	0.02	0.62	0.62	0.04	0.04
	0.9	0.03	0.02	1.03	1.04	0.06	0.06

Table 23
Confidence Band Width for Standard Error of Measurement by Population Mean Reliability and Type of Missing Data.

		Regression Imputation		Multiple Imputation		Listwise Deletion	
		SEM	SEM	SEM	SEM	SEM	SEM
Missingness	Rhoxx	Weighted	Unwted	Weighted	Unwted	Weighted	Unwted
None	0.4	0.05	0.05	0.05	0.05	0.05	0.05
	0.6	0.03	0.03	0.03	0.03	0.03	0.03
	0.8	0.02	0.02	0.02	0.02	0.02	0.02
	0.9	0.01	0.01	0.01	0.01	0.01	0.01
Random	0.4	0.04	0.04	1.15	1.15	0.08	0.08
	0.6	0.03	0.03	0.78	0.78	0.06	0.06
	0.8	0.02	0.02	0.47	0.47	0.03	0.03
	0.9	0.01	0.01	0.31	0.31	0.02	0.02
Systematic	0.4	0.04	0.04	1.00	1.00	0.08	0.08
	0.6	0.03	0.03	0.69	0.69	0.05	0.05
	0.8	0.02	0.02	0.43	0.43	0.03	0.03
	0.9	0.01	0.01	0.29	0.29	0.02	0.02

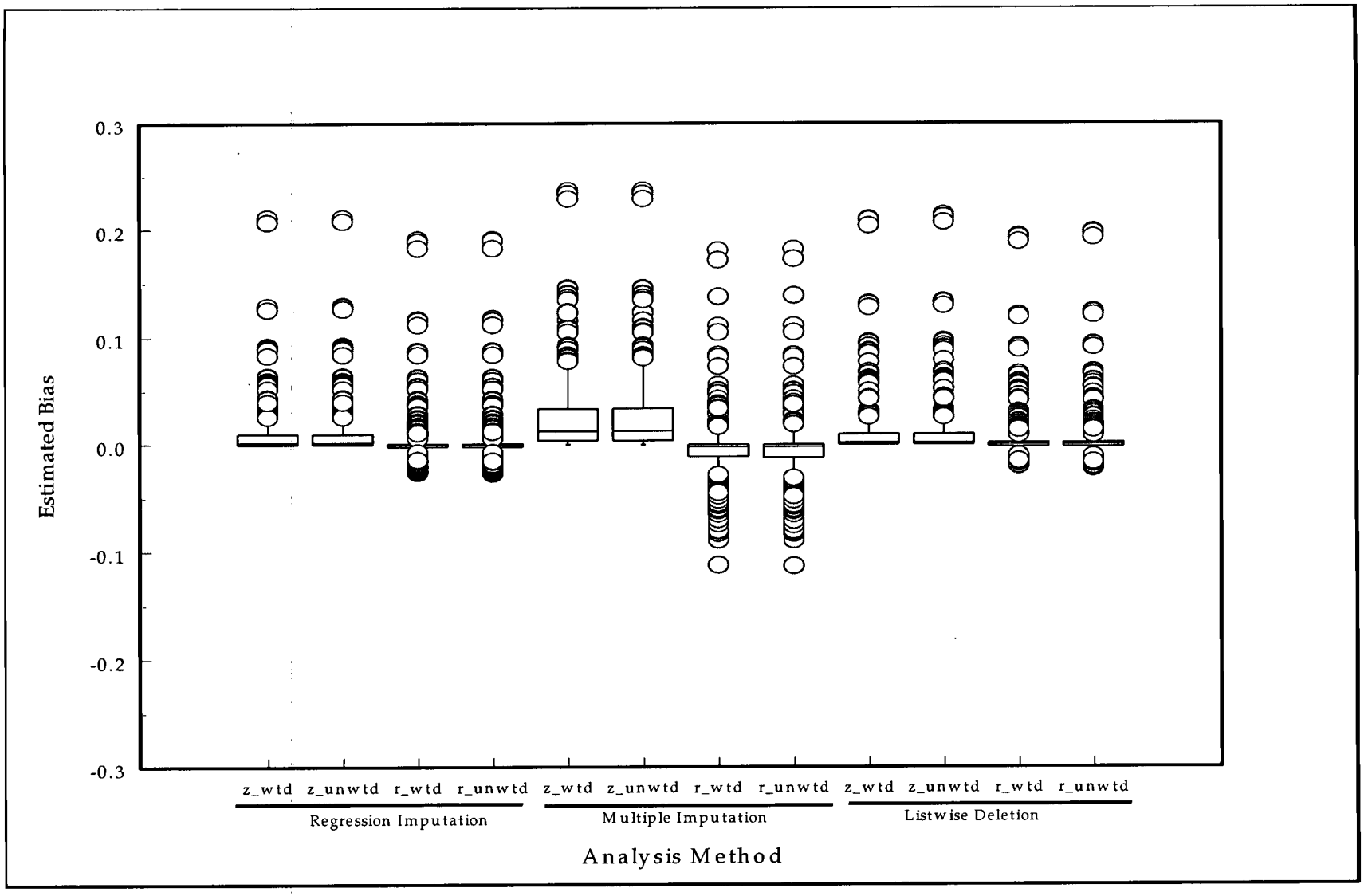

Figure 1
Distribution of Bias in the Estimation of Reliability Coefficients.

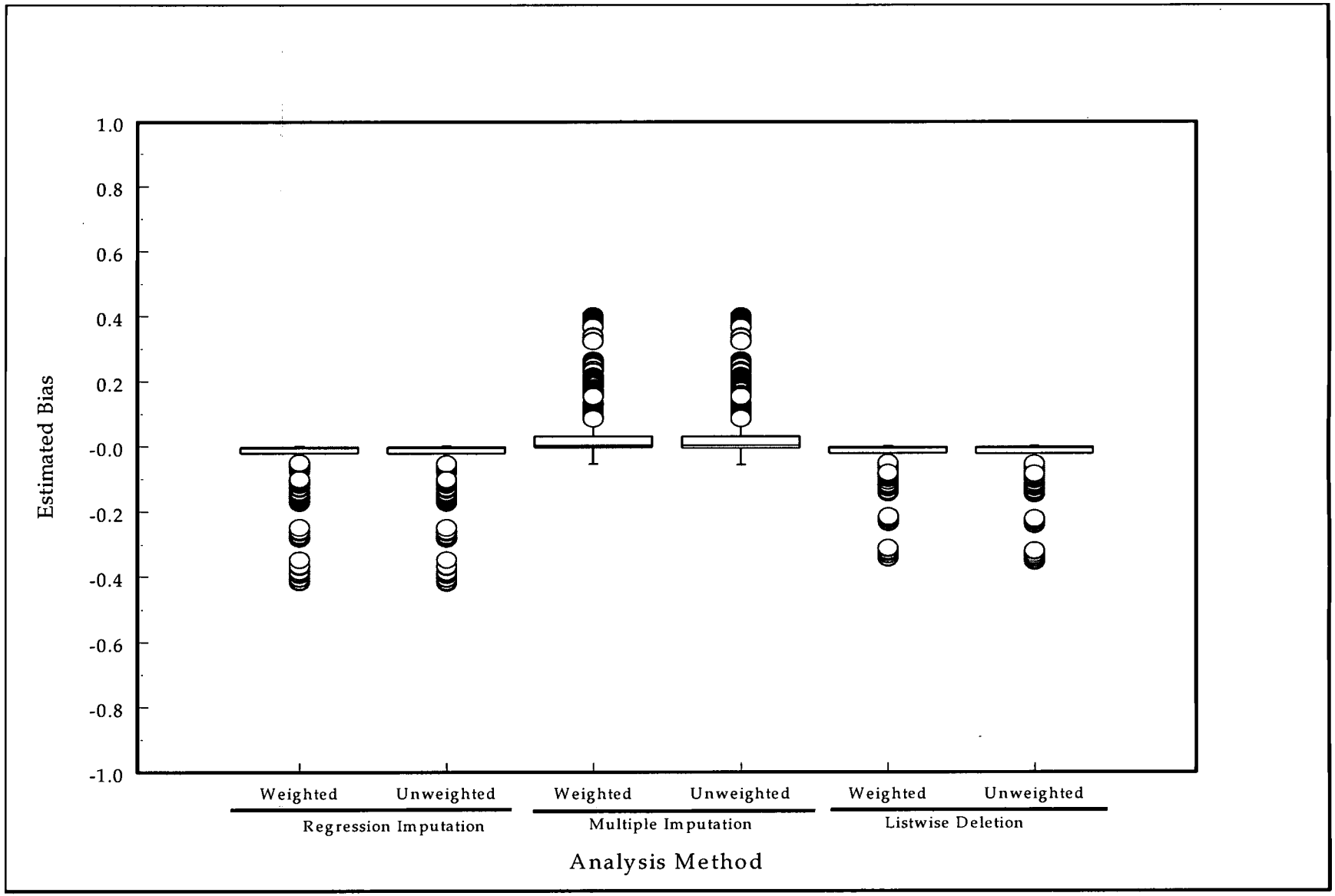

Figure 2
Distribution of Bias in the Estimation of the Standard Error of Measurement.

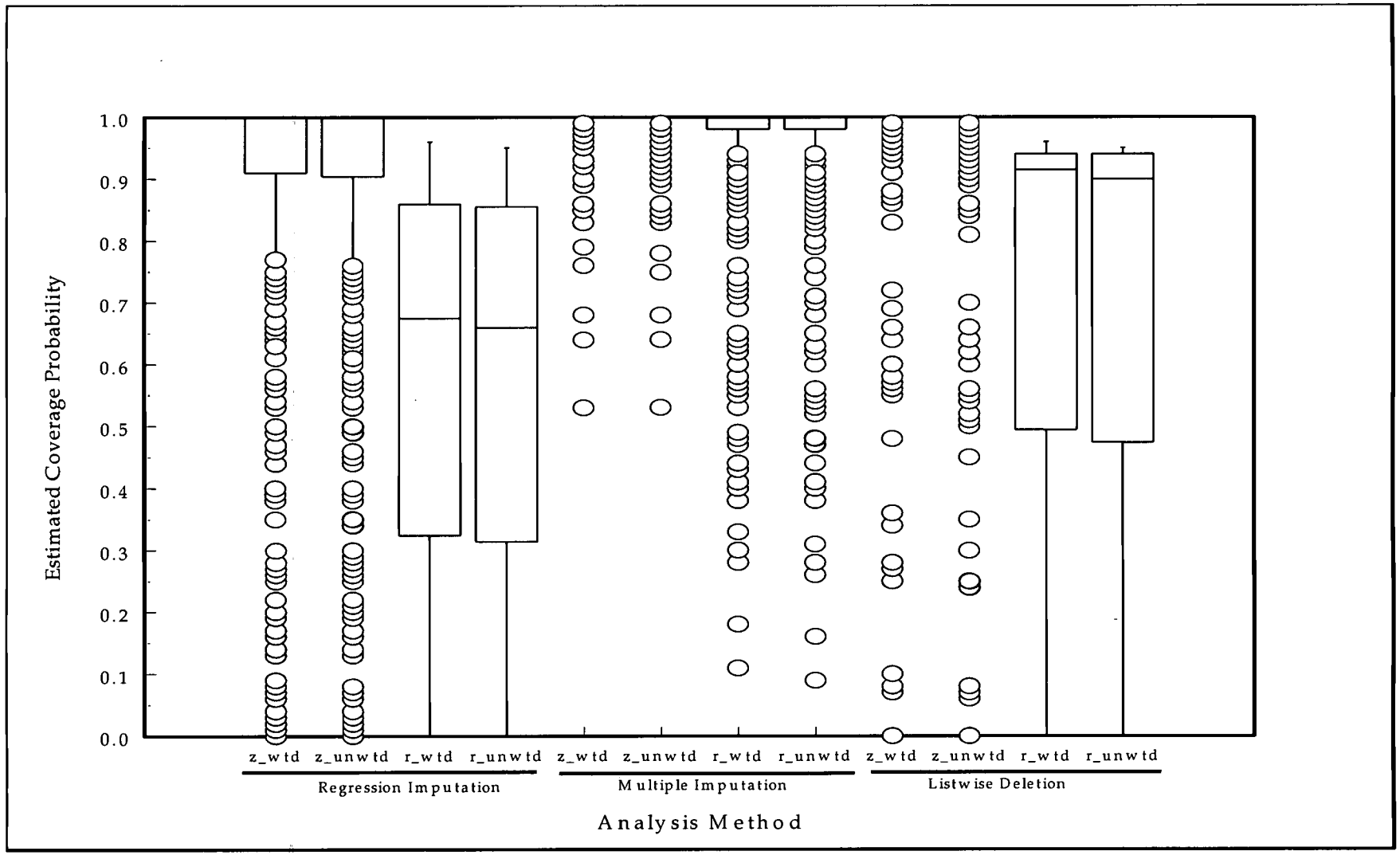

Figure 3
Distribution of Coverage Probabilities for the Estimation of Reliability Coefficients.

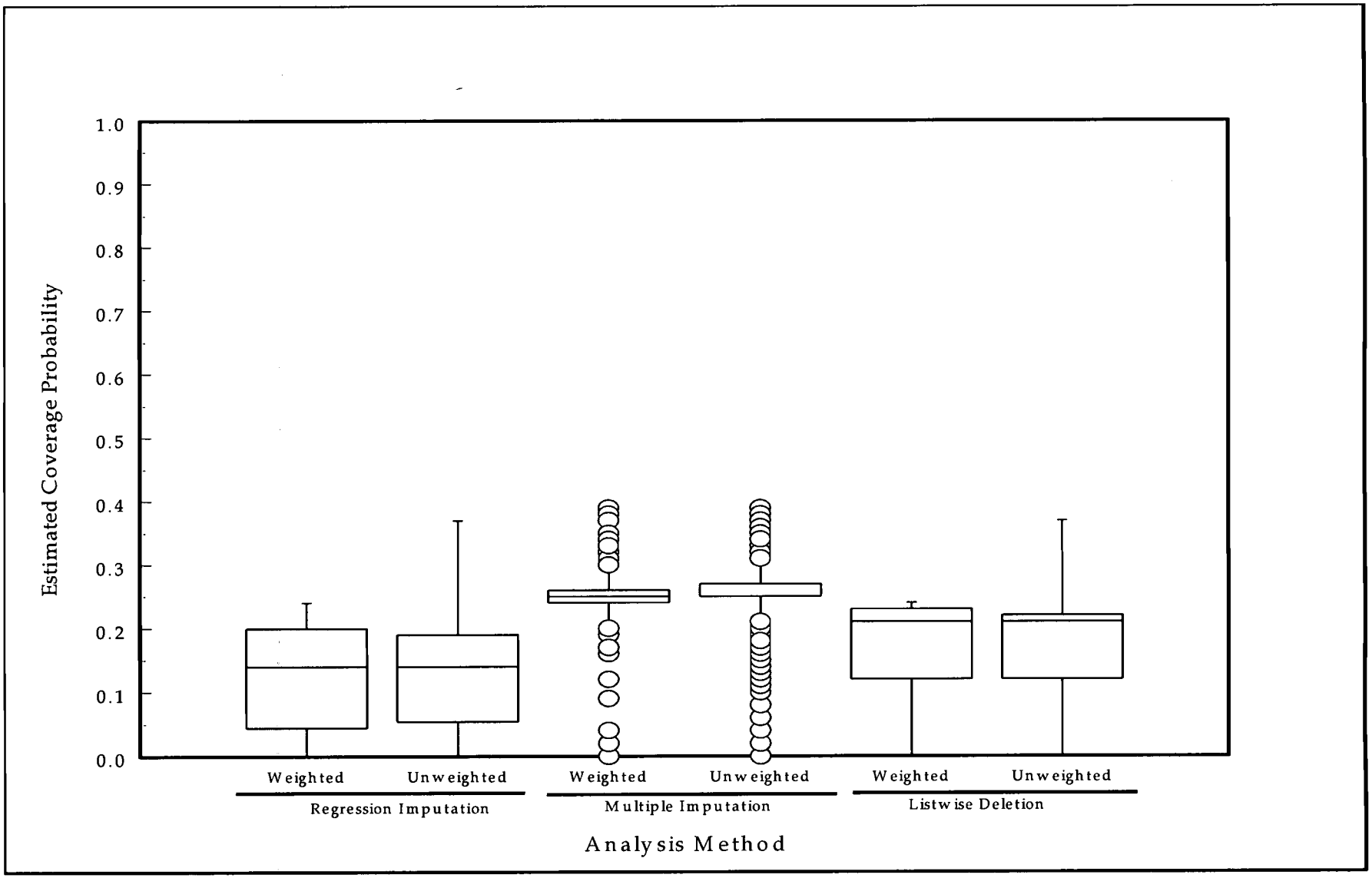

Figure 4
Distribution of Coverage Probabilities for the Estimation of the Standard Error of Measurement.

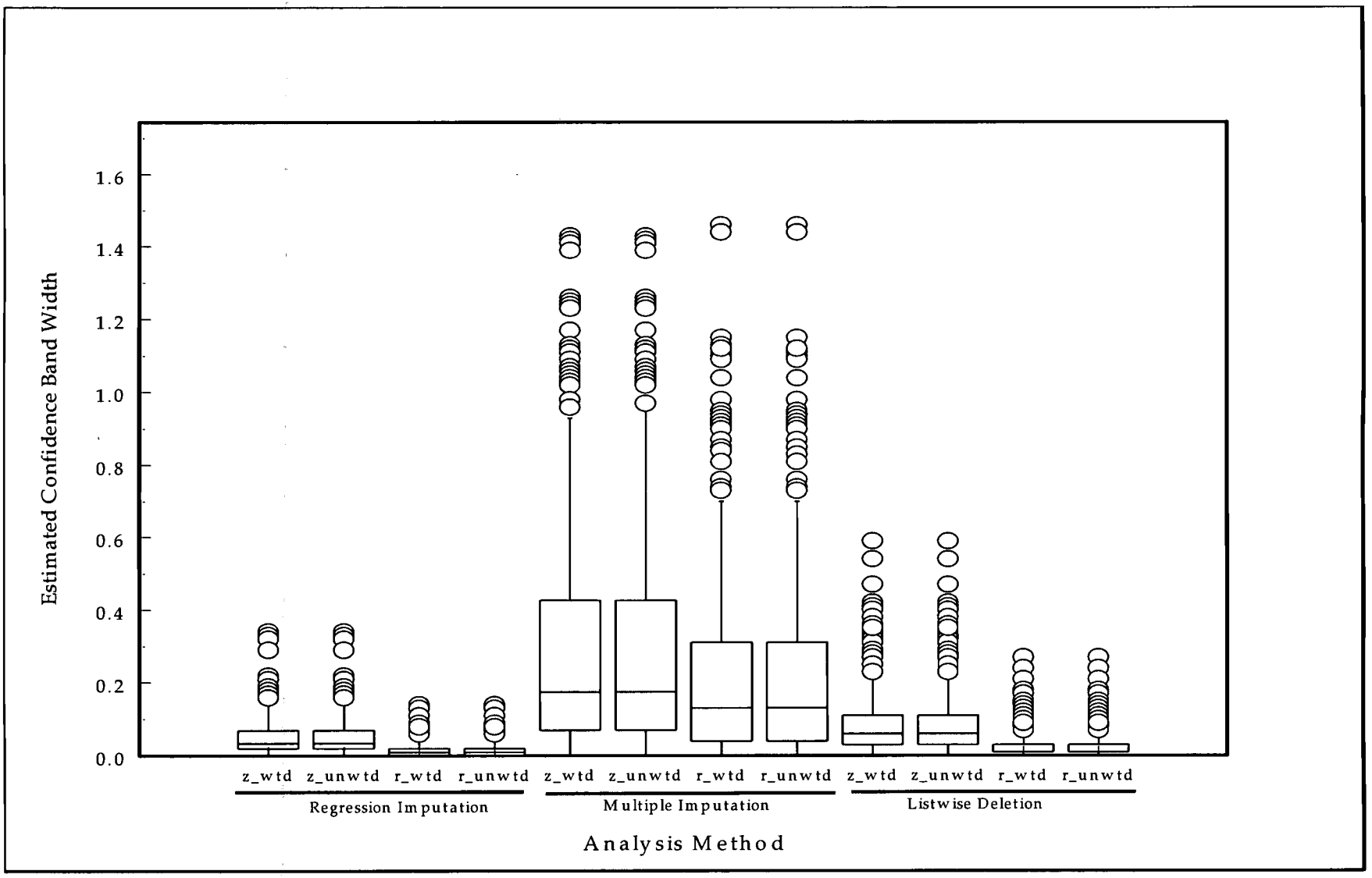

Figure 5
Distribution of Confidence Band Widths in the Estimation of Reliability Coefficients.

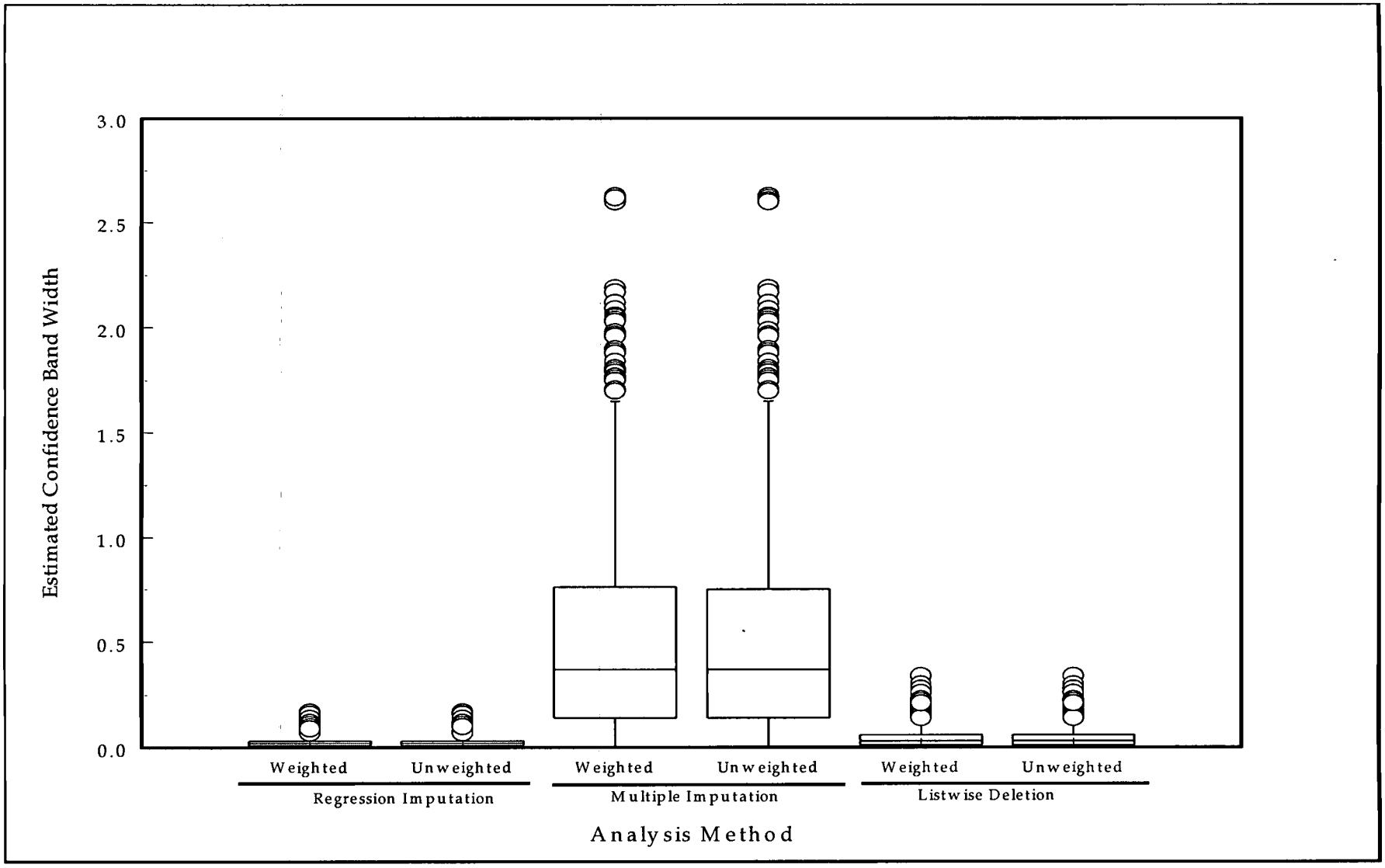

Figure 6
Distribution of Confidence Band Widths in the Estimation of the Standard Error of Measurement.

AERA

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

TM034187

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The "RG Sausage's" Missing Ingredients: Investigating the Validity of the Relibility Generalization Study Design	
Author(s): Jeanine Romano & Jeffrey D. Kromrey	
Corporate Source: University of Tampa	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: Jeanine Romano / Instructor
Organization/Address: University of Tampa	Telephone: 813 253 3333 FAX:
	E-Mail Address: JRomano@ut.edu Date: May 15, 2003

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)